Australian Government
Department of Defence
Defence Science and
Technology Organisation

# Towards an Evaluation of Air Surveillance Track Clustering Algorithms via External Cluster Quality Measures

*Matthew C. Lowry*

**Command, Control, Communications and Intelligence Division**
Defence Science and Technology Organisation

DSTO-TR-2800

## ABSTRACT

Clustering is a data mining technique for analysing large data sets and finding groups of elements within the data set that are similar to each other. The use of clustering on archives of historical air surveillance track data would enable the discovery of flights that exhibited similar behaviour and followed similar flight paths. However there are many different clustering algorithms available, so some method for selecting the best from the competing algorithms is required. Unfortunately the academic literature has yet to provide a general, comprehensive, and robust methodology for this task. Further the niche nature of the problem domain means the academic literature provides no direct assistance by way of reporting practical experience in the use of particular algorithms on air surveillance track data. This report aims to fill the gap by describing such a methodology for evaluating and choosing between competing clustering algorithms.

## RELEASE LIMITATION

*Approved for public release*

*Published by*
*Command, Control, Communications and Intelligence Division*
*DSTO Defence Science and Technology Organisation*
*PO Box 1500*
*Edinburgh South Australia 5111 Australia*

*Telephone: (08) 7389 5555*
*Fax: (08) 7389 6567*

*AR-015-522*
*January 2013*

**APPROVED FOR PUBLIC RELEASE**

# Towards an Evaluation of Air Surveillance Track Clustering Algorithms via External Cluster Quality Measures

## Executive Summary

Clustering is a data mining technique for analysing large data sets. The technique finds groups of elements within the data that are similar to each other, but different from other data elements outside the group. The use of clustering on archives of historical air surveillance track data would enable the discovery of groups of flights that followed the same flight path. This could enable improved capability in a variety fields including situational awareness, tactical air intelligence (automated behavioural prediction and anomaly detection, indicators and warnings, *etc.*), strategic air intelligence (historical analysis, capability assessment, *etc.*), and general efficiency dividends (higher performance of air surveillance and air intelligence operators, improved training and knowledge retention practices, *etc.*).

However there are many different clustering algorithms available, so before clustering can be used a method for selecting the best from the competing algorithms is required. Unfortunately the academic literature has yet to provide a general, comprehensive, and robust methodology for this task. Further the niche nature of the problem domain means the academic literature provides no direct assistance by way of reporting practical experience in the use of particular algorithms on air surveillance track data.

This report aims to fill the gap by describing a methodology for evaluating and choosing between competing clustering algorithms. Note that this report does not describe the outcome of actually performing an exhaustive evaluation and selection process. Rather, this report describes the methodology and experience from a trial of the methodology on a test data set of air surveillance track data. The experience was generally positive, in that the methodology achieved the desired outcome, however it is concluded improvements to the methodology can and should be sought.

*This page is intentionally blank*

# Author

**Matthew Clifton Lowry**
Command, Control, Communications and
Intelligence Division

*Matthew Lowry received his Ph.D. in Computer Science from The University of Adelaide in 2005 for developing a novel approach to a family of algorithms in distributed computing. In the same year he began working for DSTO in the Intelligence Analysis Discipline as a Data Mining Researcher. In this position he has been involved in an eclectic range of projects to support client organisations in the Defence Intelligence, National Security, and Counter-terrorism space. Matthew's work is focussed on algorithmic analysis and visualisation of large data sets, and his research interests include the analysis of geospatial and spatio-temporal data sets.*

*This page is intentionally blank*

# Contents

*This page is intentionally blank*

# 1. Introduction

The context for this report is capability development in the area of air surveillance. In particular, the goal is to analyse and exploit archives of historical air surveillance track data to identify groups of flights that exhibited similar behaviour and followed similar flight paths. This form of analysis could enable improved capability in a variety fields including situational awareness, tactical air intelligence (automated behavioural prediction and anomaly detection, indicators and warnings, *etc.*), strategic air intelligence (historical analysis, capability assessment, *etc.*), and general efficiency dividends (higher performance of air surveillance and air intelligence operators, improved training and knowledge retention practices, *etc.*).

*Clustering* is a data mining technique for analysing large data sets and finding groups of elements within the data set that are similar to each other. An introduction to the topic is beyond the scope of this report; the reader is directed to a popular textbook by Han and Kamber (2006) for an introduction and an article by Jain, Murty and Flynn (1999) for the most comprehensive survey of clustering algorithms in academic literature.

The suitability of clustering for exploiting air surveillance track data was explored in RPDE Task 20. The experience of that task was promising. A capability demonstration system using clustering was built and assessed by air surveillance experts, who formed the opinion that the system was successful, on a rudimentary level at least, in facilitating capability enhancements of the type mentioned previously. However, that project deliberately avoided the question of how to evaluate and choose between different clustering algorithms as the goal was to demonstrate that at least one clustering algorithm with some degree of effectiveness existed.

This report describes a methodology for evaluating and choosing between different algorithms. Note that this report does not describe the outcome of actually performing an exhaustive evaluation and selection process. Rather this report describes a methodology that would enable that selection process to be undertaken in a rigorous fashion, and experience from a trial of the methodology on a test data set of air surveillance track data.

# 2. Selecting a Clustering Algorithm

The field of data clustering has a long history, with many different algorithms being described in academic literature. As an indication, the review by Jain, Murty and Flynn (1999) contains over 200 references. Given the range of choice, a basis for selecting an algorithm must be established.

## 2.1 Fitness For Purpose

In selecting an algorithm, some general fitness-for-purpose criteria can be applied. Three broad criteria can be considered: the form of input assumed by the algorithm, form of output produced by the algorithm, and any parameters that the algorithm requires for operation.

### 2.1.1 Input Assumptions

To be suitable for the task at hand, we need an algorithm that at the very least is capable of operating over air surveillance track data.

On the whole, the operation of a clustering algorithm consists of two distinct aspects. Firstly there needs to be a *distance* or *similarity* function that can be applied to two elements from the data set to produce a non-negative real number describing how close or similar to each other those elements are. Secondly the algorithm needs to have a strategy for using that function to analyse the structure of the input data set and produce a result describing patterns within the data.

Therefore we can consider "generalist" algorithms that use a general strategy and can operate on any type of data, including air surveillance track data, provided the algorithm is furnished with a distance function suitable for this type of data. For example, the DBSCAN algorithm (Ester et al 1996) describes a general strategy based on notions of "neighbourhood" and "density" induced from any suitable distance function.

Alternatively we might consider "specialist" algorithms that describe a specific distance function for track data and a specialised strategy for exploiting that specific distance function. For example, the TRACLUS algorithm (Lee, Han & Whang 2007) is specifically designed to operate on track data and describes a specialised method for pre-processing tracks and assessing the similarity between segments of tracks.

However we can exclude specialist algorithms that make assumptions about the input data that are not met by air surveillance track data. For example the BIRCH algorithm (Zhang, Ramakrishnan & Livny 1996) is specifically intended for operation over a collection of uniformly-dimensioned real number vectors. The algorithm's strategy involves building a tree-like data structure that captures the distribution of the input data within its vector space using summary vectors generated by real vector arithmetic. Hence the BIRCH algorithm could not be used, as is, on track data.

### 2.1.2 Output Form

In general the output of a clustering algorithm is a mapping that assigns each element of the input data set to a cluster. However, there is some variety in the nature of the output from clustering algorithms. For example some algorithms will produce a *soft* or *fuzzy* clustering where a data element may be assigned to multiple clusters, potentially with a number indicating the strength of association to each cluster. Some algorithms produce a *hierarchical* clustering, in which clusters may be further grouped into clusters of clusters, which are further grouped into clusters of clusters of clusters, and so on.

Therefore when considering algorithms the suitability of form of the algorithm's output and the way that output is to be exploited should be considered. Strictly speaking it is not necessary to exclude any particular output form, since in general output formats can be transformed. For example, a fuzzy clustering output can be transformed to a hard clustering by selecting the most likely cluster for each data element. However intuition would suggest it best to avoid the complexity and assumptions in algorithms that produce output structure that is not relevant to the task at hand.

### 2.1.3 Parameter Requirements

In general, clustering algorithms will leave one or more of the parameters controlling their behaviour unspecified. Operation of the algorithm on a particular data set requires the selection of values for those parameters. This can have implications for the suitability of the algorithm to a particular task.

For example, some algorithms require specification of thresholds that are used to determine whether the underlying structure discovered in the input data is significant. For such algorithms the number of clusters in the output of the algorithm is variable and will depend on the input data as well as the parameter selection. The DBSCAN and TRACLUS algorithms mentioned previously are examples of this type of algorithm.

In contrast, some algorithms require specification as a parameter the number of clusters to be present in the output. These algorithms will always produce the specified number of clusters in the output, regardless of whether they are "true" clusters (in the sense of reflecting significant underlying structure in the input data). The well-known K-means algorithm (MacQueen 1967) is a prominent example of this type of algorithm. In the context of clustering air surveillance track data, where the number of clusters is not known *a priori*, such algorithms are unsuitable.

### 2.1.4 Assessment of Competing Algorithms

The fitness-for-purpose criteria discussed above allow the field of choice to be narrowed, but not to the extent that a single algorithm is clearly the best for clustering air surveillance track data. And even if a single algorithm were to be chosen on some basis, there remains the issue of selecting values for the algorithm's operational parameters. To optimise this selection, a method of exploring the choices and assessing their merit is required. This task is generally referred to as *cluster quality assessment*.

A comparison of different algorithms and parameters can be achieved by executing them over a test data set and applying some function to the outputs to rate their quality. The function could be a "goodness" or "quality" function, where higher numbers are interpreted as better. Alternatively the function could be a "badness" or "error" function, where higher numbers are interpreted as worse. Either way, such functions allow a comparative assessment to take place.

In general there are two distinct kinds of cluster quality measure: *internal* and *external* measures. An internal measure is one that operates on a clustering result and assesses its structure in terms of some abstract notion of the kind of structure that should be present in an ideal clustering result. An external measure of cluster quality is essentially a function that assesses the similarity between two different clusterings, and relies on some kind of "gold standard" (*i.e.* a test data set and a pre-defined clustering of that test data set produced manually by an expert). The manually produced clustering is treated as the ideal output for an algorithm to produce. Then the output of any algorithm applied to that data set can be assessed in terms of how similar it is to the ideal.

## 2.2 Clustering Quality Assessment with Internal Measures

The use of an internal measure for assessing the results obtained via different algorithms and parameter settings is attractive, if for no other reason, because it avoids the effort of handcrafting an ideal result for external measures to compare an algorithmic result against.

However we are still faced with the dilemma of selecting an internal quality measure from amongst the many that could be employed. It is easy to develop a function that applies to a clustering result and yields a real number; for example an empirical study reported by Nguyen and Rayward-Smith (2008) involved 44 distinct functions which the authors obtained by considering functions described in previous literature and generating additional functions by variation on themes.

Different internal quality measures will produce different values and rankings when they are used to assess algorithms, so there is the issue of which measure to use. There are indications on both analytical and empirical grounds that care should be taken in this regard.

### 2.2.1 Internal Measures Based on Pathological Ideals

There is the issue of what precisely a quality measure considers to be the ideal for a clustering result. Consider, for example, the *sum of squared distances from the centroid* measure. Given $k$ clusters denoted $C_1$, $C_2$, ..., $C_k$, with $c_i$ denoting the centroid of cluster $C_i$, and $\delta$ denoting a distance function, then the measure is:

$$\sum_{i=1}^{k} \sum_{x \in C_k} \delta^2(c_i, x)$$

That is, for each cluster, compute the square of the distance between each element of the cluster and the cluster's centroid, and take the sum of all these squared distances. This is an error style measure; zero is the minimum possible value and higher values indicate lower quality. Intuitively, this seems to be generally suitable as a cluster quality measure.

However consider the behaviour of this measure *in extremis*. First, consider a hypothetical perfect algorithm that is able to produce output indistinguishable from what a human expert deems perfect. Even in this circumstance, the measure will assign a non-zero value to the clustering. In other words this measure will describe a clustering result as having some degree of error even if a human expert would describe that result as perfect.

Further, consider the nature of a clustering result that this measure would actually describe as having no error. Such a result will necessarily have the same number of clusters as there are distinct data elements of the input data; each data element assigned to a separate cluster containing only itself. So the ideal, perfect result according to this quality measure is the clustering in which no structure is reported by the algorithm. The perfect algorithm would be one that does no work and essentially returns the input data set as the output result.

Note that the well-known K-means algorithm is explicitly designed to find a solution that minimises this very measure. Yet K-means has been in use for over four decades and remains popular. This apparent paradox is resolved when one considers the fact that K-means requires the number of clusters in its output to be specified as a parameter. So the algorithm is constructed in a way that constrains the solutions it will consider, and this prevents the search for solutions from descending towards the pathological result that attempting to minimise the measure would otherwise encourage.

### 2.2.2 Empirical Evidence Regarding Internal Measures

Nguyen and Rayward-Smith (2008) report an experiment in which a variety of cluster results were ranked with a variety of internal cluster quality measures. They found that the different quality measures were not always in agreement, and in particular the result given first rank by some particular measure was typically not given first rank by the other measures. Also, they found that some measures gave first rank to semi-synthetic results that the researchers had deliberately constructed to be sub-optimal. Similar experience was reported by Raskutti and Leckie (1999).

### 2.2.3 Corollary

The foregoing discussion leads to the conclusion that it would be dangerous to naïvely choose some internal cluster quality measure that intuitively seems an apt measure, and proceed to evaluate algorithms solely on the basis of that measure. For any given analytical goal and data domain, before the quality of the output of competing clustering algorithms can be assessed with internal cluster quality measures, an empirical study will be needed to select the measure that is best suited for assessing the competing algorithms.

Such an empirical study would be of the form described by Nguyen and Rayward-Smith (2008). It would necessarily involve the generation of a collection of clusterings with known quality, so that the ranking given by an internal quality measure can be compared for consistency with the known true ranking. A minimal procedure for obtaining such a collection of clusterings is to have a human expert manually construct the ideal clustering, and then automatically generate a sequence of suboptimal clusterings by progressive random permutation of the ideal.

## 2.3 Clustering Quality Assessment with External Measures

If one is resigned to the effort of manually constructing a gold-standard clustering for some test data set as a necessary part of any procedure to select between competing clustering algorithms, then it would be sensible to simply use that gold-standard clustering to directly assess the quality of an algorithm's output. As noted before, an external cluster quality measure is a binary function that takes an ideal clustering as one parameter and an algorithmically generated clustering as the other parameter, and returns a number to characterise the similarity (or lack thereof) between the ideal and algorithmically generated clusterings. Once the effort of constructing the hand-crafted ideal clustering has been made, selecting from competing clustering algorithms is straight-forward; simply select the algorithm that, when applied to the same test data, produces output most like the gold-standard.

Note that we still have the dilemma of selecting the external quality measure that does the best job of assessing the similarity between clusterings. However this report proposes a methodology for achieving this. The methodology requires some prior experience; specifically experience with at least one clustering algorithm that produces output that is judged to be good in the sense of being similar to the ideal. Given this basis, the methodology allows selection of an external quality measure suitable for the type of data and analysis being performed, and has the convenient side-effect of "boot-strapping" the subsequent task of using the measure to choose between competing clustering algorithms.

The methodology is motivated by the intuition that the measure should readily discriminate good clusterings from bad clusterings, and that a random clustering will almost certainly be a bad clustering. If an external quality measure fails to recognise that random clusterings are generally quite bad, then it does not have a useful degree of *discriminatory power*. This is depicted in Figure 1.

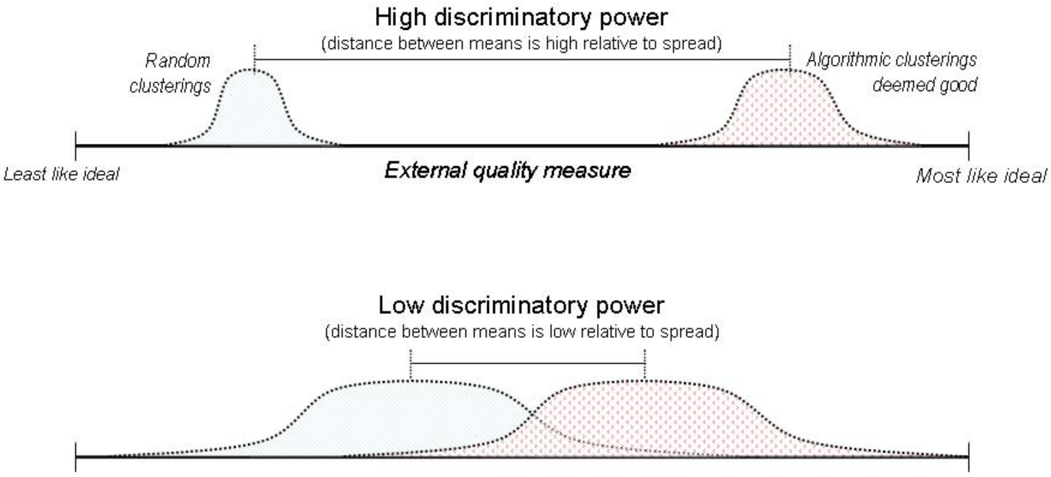
High discriminatory power
(distance between means is high relative to spread)
Random clusterings
Algorithmic clusterings deemed good
Least like ideal
External quality measure
Most like ideal
Low discriminatory power
(distance between means is low relative to spread)

*Figure 1: Discriminatory power of an external cluster quality measure*

More concretely, if the measure is applied to a collection of good clusterings, and also applied to a collection of random clusterings, then the distribution of the measure across these two collections should be well separated.

Let $M$ denote the measure. Let $I$ denote the ideal clustering.

Let $G_1, G_2, ..., G_n$ denote a collection of good clusterings.

Let $R_1, R_2, ..., R_n$ denote an equal sized collection of random clusterings with broadly similar structure, in the sense of having a similar number of clusters and distribution of elements across the clusters as the good clusterings.

Let $G_M$ denote the set $\{M(I,G_1), ..., M(I,G_n)\}$ and let $R_M$ denote the set $\{M(I,R_1), ..., M(I,R_n)\}$.

Then let the discriminatory power of the measure be given by

$$\frac{|\,\mu(G_M) - \mu(R_M)\,|}{\sqrt{s(G_M)\,s(R_M)}}$$

where $\mu$ denotes arithmetic mean and $s$ denotes sample standard deviation.

Based on this conception of discriminatory power, the following methodology can be used for first selecting the external quality measure with the best discriminatory power, and then using that measure to select amongst competing clustering algorithms.

1. Choose a test data set that is generally representative of the type of data in which clustering will be performed.
2. Manually construct an ideal clustering on the test data.
3. Obtain a set of algorithmic clusterings by applying to the test data any number of clustering algorithms and parameter settings that are informally judged by an expert to be producing a good approximation of the ideal.
4. Generate the same number of random clusterings, in a manner that produces clusterings with a broadly similar distribution of elements (in terms of the number of clusters and the number of elements in clusters).
5. For all external quality measures under consideration, evaluate the measure's discriminatory power using the clusterings obtained in steps 3 and 4.
6. Optionally, obtain the output of any additional clustering algorithms and parameter settings deemed worthy of evaluation.
7. Select the clustering algorithm and parameter setting that produces output most similar to the ideal, as assessed by the measure with the highest discriminatory power.

A trial of this methodology using air surveillance track data is reported subsequently in Section 3.

## 2.4 Related Work

Scholarly articles that address the same area as this report are rare.

It is typical of the literature that when a new clustering algorithm or improvement on an existing algorithm is proposed, the justification is based on improved efficiency (in terms of time and space requirements for the computation). When the output quality is addressed at all, typically an informal argument that the output will be good is made on the basis of the algorithm's construction and its performance on nugatory data sets with known gold-standards such as the famous Fisher iris data set (Fisher 1936). The articles that do go beyond this basic level of quantitative analysis typically involve contextual factors that limit the relevance to the current topic of air surveillance track data.

For example, Schmitzer-Torbert *et al* (2005) describe two internal cluster quality measures applied in the context of neurophysical experiments; specifically analysis of neuronal excitation recordings from multi-channel electrodes to discriminate noise from "true" voltage spikes. The authors claim the two internal cluster quality measures they assessed were good for this purpose because the measures were consistent with the judgement of a human expert. They are motivated by the same fundamental issue of finding a cluster quality function that is consistent with the judgement of a human expert so it can be used to assess clustering algorithm output. However, they do so in a context far removed from the context of this report; in particular, the two specific internal clustering quality measures they advocate cannot be applied to air surveillance track data.

Meilă (2005) addresses the issue of comparative assessment of the merit of different external quality measures, but does so from a highly abstracted axiomatic perspective.

Vinh, Epps & Bailey (2009) describe a quantitative analysis of three external quality measures. Specifically they consider measures adjusted for chance agreement and compare the well-known Adjusted Rand Index with two other measures adjusted in the same manner. However their motivation is quite different from the present context, and aspects of their methodology limit its general applicability. Their methodology revolves around a carefully constructed synthetic test data set in which the number of true clusters is known, and using a specific clustering algorithm in which the number of clusters in the output is fixed by parameter to the algorithm. In addition their motivation and conclusions are tangential to this report; they are focussed on the nature of the adjustment for chance and delineating the general character of data sets for which the adjusted measures perform well.

The literature does not appear to have received a comprehensive survey of the various measures that have been proposed. Some authors present brief and non-exhaustive surveys in the context of proposing new methodologies. Meilă (2007) provides a discussion of many previous approaches in the context of describing a new approach, and similarly Ben-Hur, Elisseeff & Guyon (2002) discuss a number of previous approaches by way of presenting a generalisation of those approaches.

# 3. Experiments

To make use of the methodology outlined in Section 2.3, a representative test data set must be selected. Section 3.1 describes the air surveillance track data set that was obtained for the experiment. The clusterings generated on this data for the experiment are described in Section 3.2. The external quality measures that were selected and used in the experiment are described in Section 3.3. The results and discussion thereof are given in Sections 3.4 and 3.5 respectively.

## 3.1 Test Data

The test data set is a collection of 5000 tracks, where each track is a sequence of fixes and each fix specifies a timestamp and a geodetic location (*i.e.* latitude and longitude relative within some geodetic coordinate system). Each track in the data set describes the real-world movement of an aircraft. The tracks were taken from the *Recognised Air Picture* (RAP) generated by RAAF 41 Wing. More precisely, the tracks were taken from an unclassified subset of the RAP. The subset covers the air space in and around Australia's Flight Information Region, but fixes and tracks corresponding to certain time periods, geographic regions, and flights are eliminated as part of a classification downgrading procedure. The data covers a period from approximately 2008-04-21 00:00 UTC to 2008-04-22 12:00 UTC.

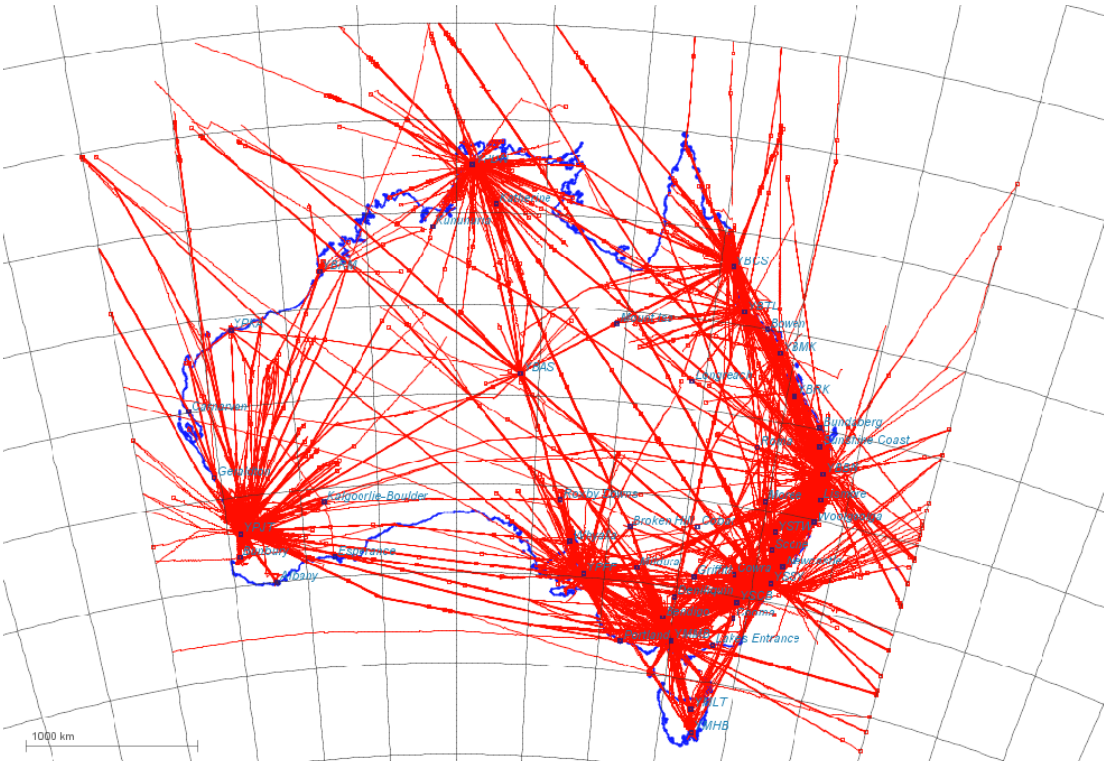

*Figure 2: Tracks of the test data set*

Figure 2 shows a rendering of the test data set. The coastline of Australia is rendered in blue, and the tracks of the test data set are rendered in red. Note that individual tracks are rendered with thin lines - where there are numerous tracks following a similar flight path their rendering merges to form the thicker lines visible on the map.

### 3.1.1 Data characteristics, filtering, and post-processing

The RAP is a real-time product generated by collating and fusing data from numerous sources and sensors, including civilian and military air traffic control systems and radars. When an aircraft is flying within sensor coverage, ideally its position will be reported with a track that is updated at a rate of 5 fixes per minute. However, due to its nature as a real-time data stream, the RAP unavoidably contains gaps and errors from numerous sources, such as noise and transient failures in sensors. Further, the logging system that was used to record the RAP data stream was itself subject to transient failures and data loss. The result was a raw data set with numerous errors including short tracks that are spurious (*i.e.* noise), single flights broken into multiple disjoint tracks in the data, and multiple separate flights by a single aircraft concatenated into a single track in the data.

To address this, the raw data was post-processed to obtain the test data set. Tracks less than 30 minutes in duration or less than 200 kilometres in total distance travelled were discarded. Tracks more than 5 hours in duration or more than 5000 kilometres in total distance travelled were also discarded. Further, tracks were discarded if they did not either start or finish in the vicinity of a major aerodrome (defined as the 30 busiest aerodromes in Australian territory, ranked by total aircraft movements, as compiled by Airservices Australia for 2010). The result of this filtering was a test data set dominated by routine passenger transport flights, including domestic flights between major cities, international flights to and from major cities, and also flights between major cities and remote locations (*e.g.* charter flights for "fly-in fly-out" workers at remote mine sites). However many other types of flights can be found in the data set; *e.g.* student pilots engaged in flying exercises, Coastwatch surveillance flights, police and medical transport helicopters, *etc.* The mean duration of the tracks is 83 minutes and the mean total distance travelled is 847 kilometres.

Another post-processing action was to compress the raw track data using the Douglas-Peucker algorithm (Douglas & Peucker 1973). This algorithm takes a sequence of points in a track and removes points such that the error caused by the removal (i.e. the distance between the original track and simplified version) is no greater than a given threshold. The aim was to reduce the computational expense for functions that measure track separation (these functions typically have a time complexity that is at least linear in the number of fixes in the two tracks). For pre-processing of the test data set a threshold of 250 metres maximum positional error was used. This led to an average compression rate of 90% for the tracks.

A result of the simplification process is to change the distribution of the track update rates (*i.e.* the rate at which fixes are given in a track, averaged over the course of the track). The tracks that exhibit more straight flying can generally have more fixes removed without introducing significant error into the track. In the simplified test data, the mean track update rate is 1.2 fixes per minute (*i.e.* approximately 50 seconds between fixes). There is however variation in this rate amongst the tracks; the least rapid is 0.03 fixes per minute and the most rapid is 3 fixes per minute. For a given track the update rate can vary significantly over the course of the track.

## 3.2 Test Clusterings

This section describes the clusterings on the test data that were used in the experimental procedure. To follow the methodology proposed in Section 2.3, three types of clustering must be obtained. Firstly an ideal or "gold-standard" clustering must be manually generated by a human expert. Secondly one or more algorithmic clusterings that are known to be good approximations of the ideal must be generated. Thirdly a collection of randomly generated clusterings that are bad approximations of the ideal must be generated.

### 3.2.1 Hand-crafted Clustering

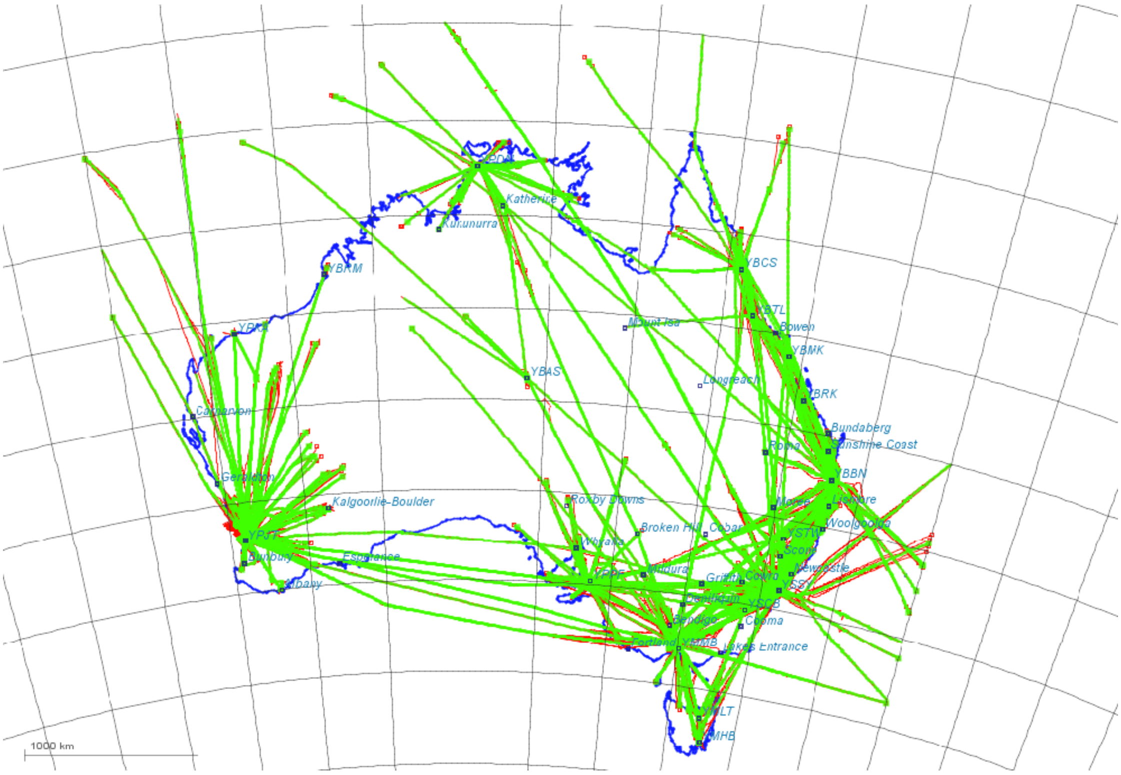

*Figure 3: Manual clustering of the test data*

The manually constructed clustering is shown in Figure 3. The clusters are depicted with red and green lines. The green lines depict the representative track of each cluster; *i.e.* the artificial track generated by averaging the component tracks in the cluster.

The clustering was constructed using the basic criteria of at least three tracks starting and finishing at the same location, following the same approximate flight path, and travelling at the same approximate speed. The tracks that could not be assigned to a cluster in accordance with these criteria were marked as outliers, and are not depicted in Figure 3. The result was a clustering containing 415 clusters and 2030 outliers. The largest cluster contained 96 tracks, although most clusters were much smaller; the mean cluster size is 7.2 tracks per cluster, and the median cluster size is only 4 tracks per cluster.

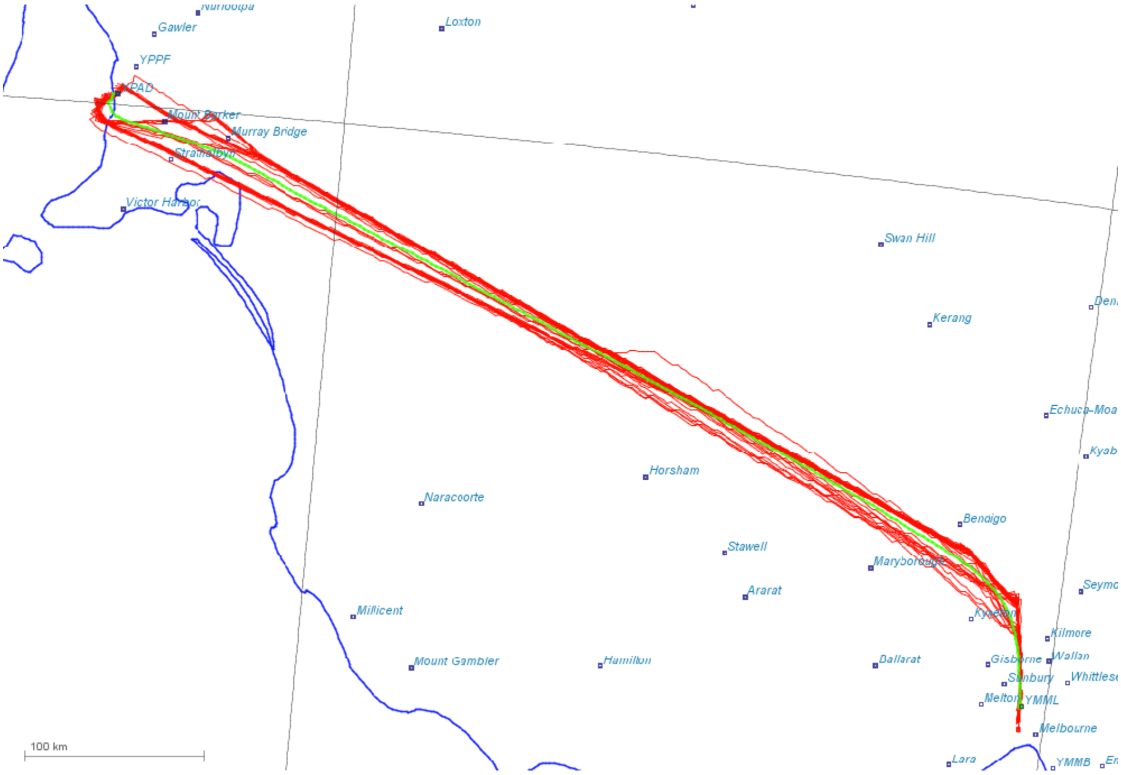

*Figure 4: Example cluster from manual clustering*

An example of a single cluster from the manual clustering is shown in Figure 4. This example is a cluster of routine passenger flights from Adelaide to Melbourne. As can be seen from the figure, there is some variation in the flight path amongst the tracks. However this is minor variation due to differences in prevailing wind and air traffic control directions received by different flights. The length of the flight path in this example cluster is 680 kilometres and the average flight time is 70 minutes, which corresponds to the expected behaviour for routine passenger transport flights made by turbofan-powered aircraft. A track in which the same flight path was flown at a slower speed (as would be the case with a turboprop-powered aircraft, for example) would not be included in the same cluster.

### 3.2.2 Agglomerative Algorithmic Clusterings

A set of clusterings were constructed algorithmically using a simple eager agglomerative strategy, two different track similarity functions, and a variety of parameter settings. The details are not particularly relevant in the present context so only a brief description is given below.

Treat the tracks from the data as a queue of unprocessed tracks. Take the next track from the queue; call this track the "seed" track. Search the remaining content of the queue for other tracks that are close to the seed track. If at least 2 such other tracks are found, remove those tracks from the queue and form a cluster with the seed track. Otherwise, declare the seed track an outlier. Repeat until the queue is empty.

Within this strategy, two different track similarity functions were employed. The first function was simply the mean separation between two tracks, sampled over time. That is, take the separation between the position of the tracks at 1 minute after the start of each track, after 2 minutes, et cetera, and take the mean of these separations. Previous experience had revealed this track similarity function combined with the eager agglomerative strategy tends to produce good output. This function is referred to as `Mean` in the following table.

The other track similarity function also computes mean separation, but adds two refinements. Firstly, it adds a length mismatch penalty which increases the apparent separation between tracks by a factor of the difference in duration between the tracks. Secondly, a rolling window is used and the peak mean value as the window rolls over the duration of the tracks is used instead of computing the mean over the entire duration. Previous experience had suggest these refinements improve the output. This function is referred to as `PRM` (for "Peak Rolling Mean") in the following table.

By combining these track separation measures with the algorithm strategy, clusterings were generated at six different separation thresholds. The basic statistics of the clusterings produced are listed in Table 1.

*Table 1: Algorithmic clusterings: parameters and basic statistics*

| ID | Separation Measure | Measure Parameters | Distance Threshold | Num. Clusters | Mean Cluster Size | Median Cluster Size | Num. Outliers |
|---|---|---|---|---|---|---|---|
| Mean-30k | Simple Mean | n/a | 30 km | 337 | 5.90 | 4 | 3011 |
| Mean-40k | | | 40 km | 416 | 8.63 | 5 | 1410 |
| Mean-50k | | | 50 km | 420 | 9.39 | 5 | 1055 |
| Mean-60k | | | 60 km | 403 | 10.28 | 5 | 858 |
| Mean-70k | | | 70 km | 369 | 11.62 | 6 | 712 |
| Mean-80k | | | 80 km | 350 | 12.54 | 6 | 611 |
| PRM:50-30k | Peak Rolling Mean | Rolling window size = 15 mins. Length mismatch penalty = 50 metres/second. | 30 km | 427 | 7.49 | 4 | 1801 |
| PRM:50-40k | | | 40 km | 375 | 6.79 | 4 | 2454 |
| PRM:50-50k | | | 50 km | 405 | 7.06 | 4 | 2139 |
| PRM:50-60k | | | 60 km | 412 | 7.46 | 4 | 1928 |
| PRM:50-70k | | | 70 km | 427 | 7.76 | 4 | 1685 |
| PRM:50-80k | | | 80 km | 427 | 8.18 | 5 | 1507 |
| PRM:250-30k | Peak Rolling Mean | Rolling window size = 15 mins. Length mismatch penalty = 250 metres/second. | 30 km | 337 | 5.90 | 4 | 3011 |
| PRM:250-40k | | | 40 km | 364 | 6.48 | 4 | 2643 |
| PRM:250-50k | | | 50 km | 400 | 6.71 | 4 | 2318 |
| PRM:250-60k | | | 60 km | 404 | 7.17 | 4 | 2125 |
| PRM:250-70k | | | 70 km | 412 | 7.47 | 4 | 1922 |
| PRM:250-80k | | | 80 km | 427 | 8.18 | 5 | 1507 |

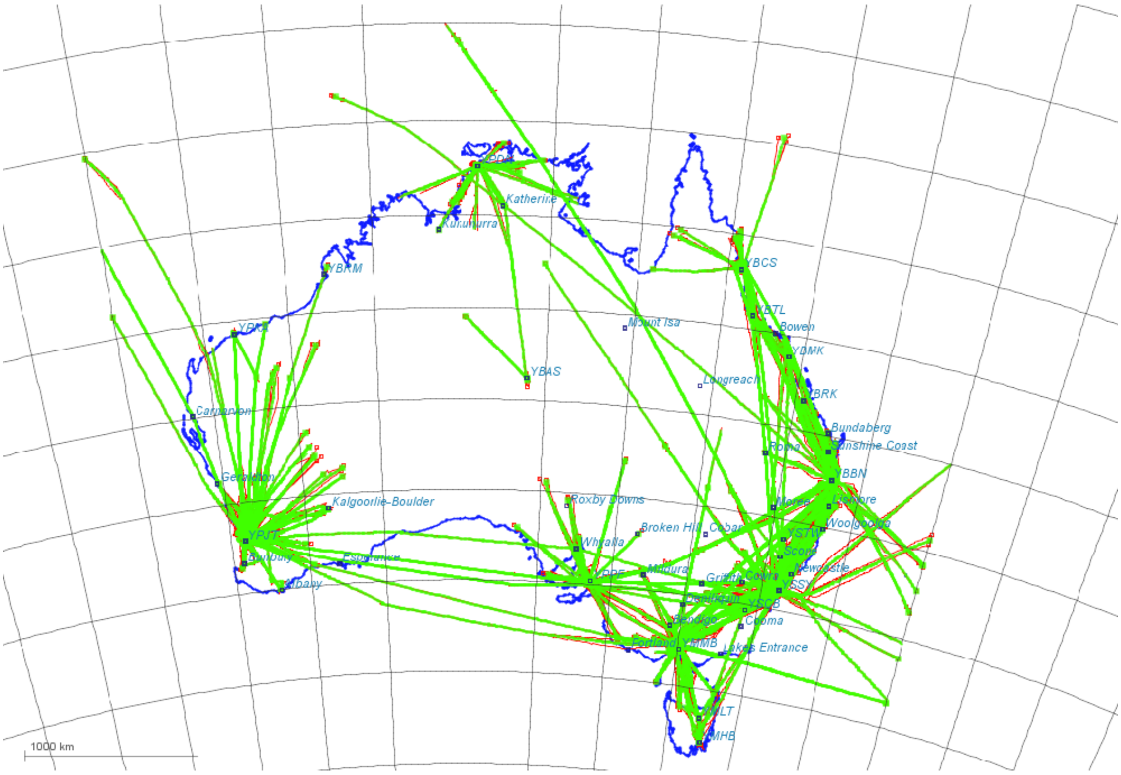

*Figure 5: Example of an algorithmic clustering of the test data*

An example of the algorithmic clustering results is shown in Figure 5. In particular, the figure is of the result `PRM:50-50k`, and as with the previous figures the outlier tracks are not shown. The results appear consistent with the manual clustering. For example, the cluster of Adelaide to Melbourne flights found in the manual clustering and depicted in Figure 4 has a counterpart discovered by the algorithm, depicted below in Figure 6. The algorithm has succeeded in finding a cluster that appears to be the same as the manually created cluster, however there are some small differences between the clusters which are more apparent on closer examination. Also there are other discrepancies between the clusterings; *e.g.* some small clusters in the ideal are missed entirely by the algorithm.

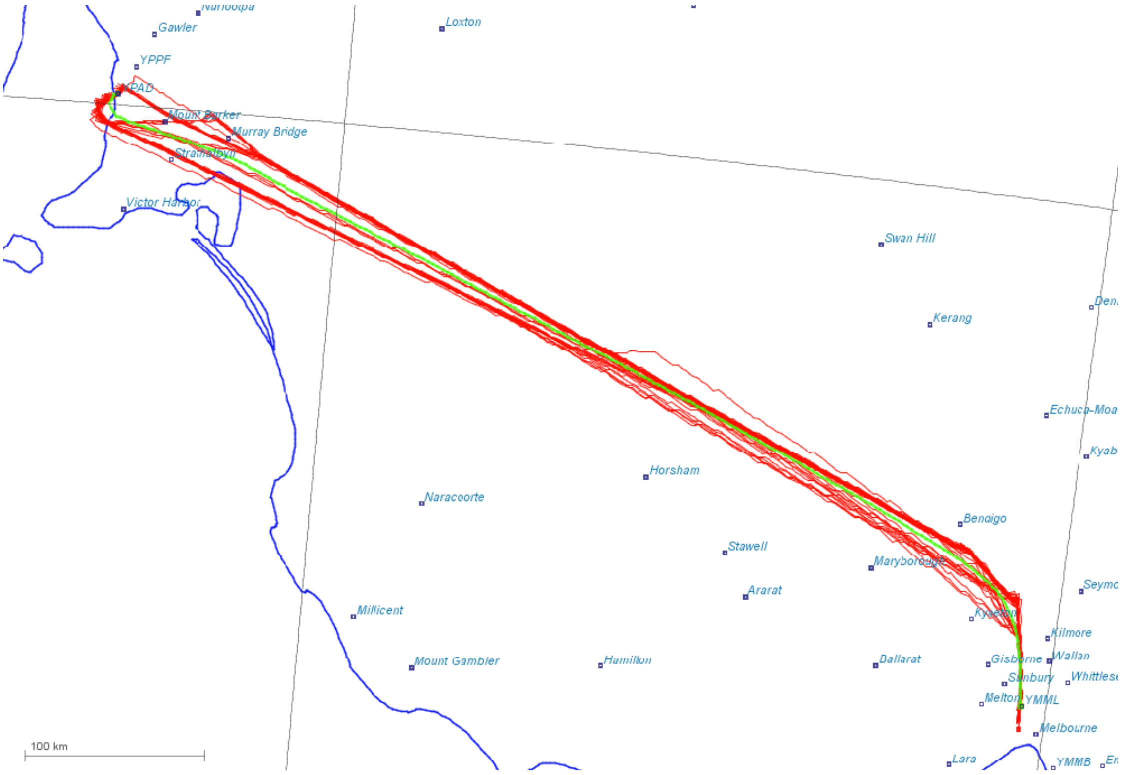

*Figure 6: Example cluster from an algorithmic clustering; good match to manual cluster in Figure 4*

It is important to note that the examples in Figures Figure 5 and Figure 6 were deliberately selected because they were generated with parameter settings that induced the algorithm to produce clusters similar to the manual clustering. Other results from less favourable parameter settings contained lower quality clusters; for example see Figure 7 which depicts a cluster from result `Mean-80k`. Because of the high threshold for considering tracks sufficiently close to form a cluster, the algorithm has formed a single cluster from flights between Brisbane and five distinct regional towns. In the manual clustering, these flights are placed into separate clusters.

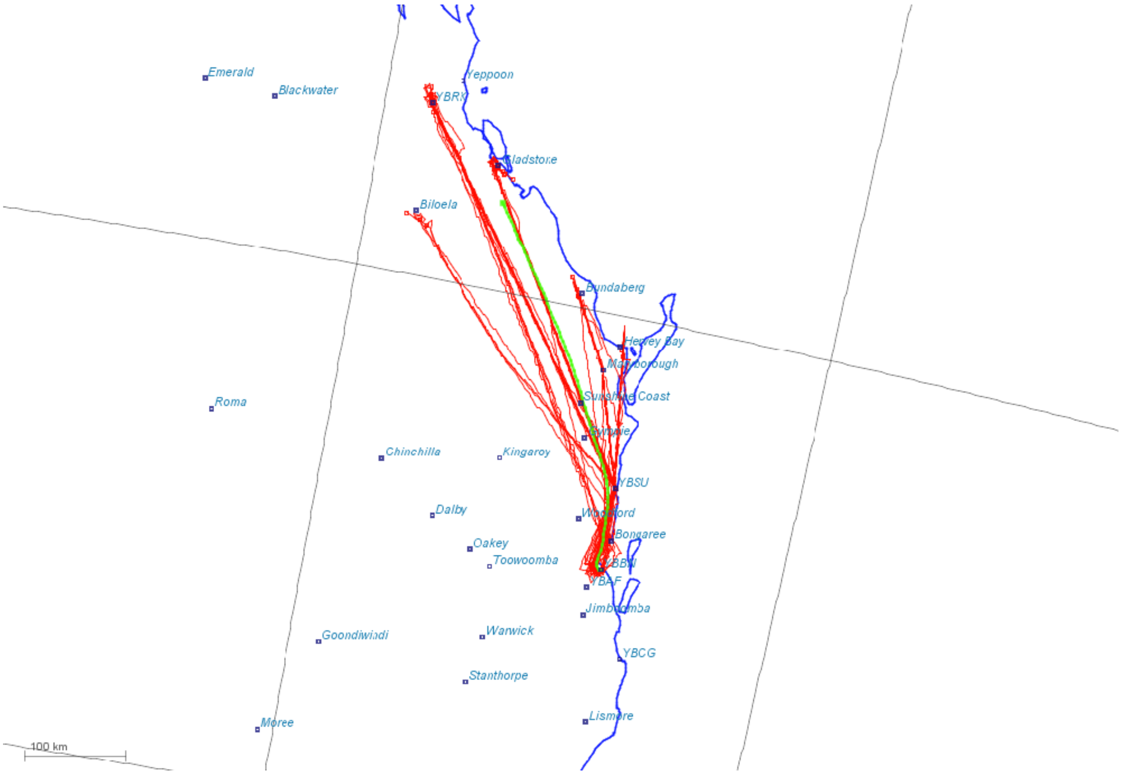

*Figure 7: Example cluster in an algorithmic clustering; matches a merging of multiple manual clusters*

### 3.2.3 Random Clusterings

A set of clusterings were also constructed by random allocation. The allocation was controlled by two parameters - the fraction of outliers *p* and the number of clusters *k*. Each track is either assigned to the outliers with probability *p* or to clusters with probability *1-p*, and the non-outliers tracks are assigned to clusters with uniform probability.

*Table 2: Random clusterings: parameters and basic statistics*

| ID | Expected Num. Clusters | Expected % Outliers | Actual Num. Clusters | Mean Cluster Size | Median Cluster Size | Actual Num. Outliers |
|---|---|---|---|---|---|---|
| Rand-300-10 | 300 | 10% | 300 | 14.96 | 14 | 513 |
| Rand-300-20 | | 20% | 300 | 13.35 | 13 | 994 |
| Rand-300-30 | | 30% | 300 | 11.62 | 11 | 1515 |
| Rand-300-40 | | 40% | 300 | 10.27 | 10 | 1920 |
| Rand-300-50 | | 50% | 300 | 8.17 | 8 | 2550 |
| Rand-300-60 | | 60% | 299 | 6.60 | 6 | 3026 |
| Rand-400-10 | 400 | 10% | 400 | 11.27 | 11 | 493 |
| Rand-400-20 | | 20% | 400 | 10.00 | 10 | 1001 |
| Rand-400-30 | | 30% | 400 | 8.81 | 9 | 1475 |
| Rand-400-40 | | 40% | 400 | 7.46 | 7 | 2018 |
| Rand-400-50 | | 50% | 400 | 6.23 | 6 | 2507 |
| Rand-400-60 | | 60% | 397 | 4.99 | 5 | 3020 |
| Rand-400-10 | 500 | 10% | 500 | 9.02 | 9 | 486 |
| Rand-500-20 | | 20% | 500 | 8.01 | 8 | 995 |
| Rand-500-30 | | 30% | 500 | 7.02 | 7 | 1488 |

| ID | Expected Num. Clusters | Expected % Outliers | Actual Num. Clusters | Mean Cluster Size | Median Cluster Size | Actual Num. Outliers |
|---|---|---|---|---|---|---|
| Rand-500-40 | | 40% | 500 | 5.96 | 6 | 2020 |
| Rand-500-50 | | 50% | 496 | 5.03 | 5 | 2503 |
| Rand-500-60 | | 60% | 490 | 4.19 | 4 | 2949 |

The basic statistics of the random clusterings generated are listed in Table 2. Note that the actual number of clusters and outliers varies slightly from the expected number because of the random generation. In particular, if a cluster had no tracks assigned to it then it was omitted from the result.

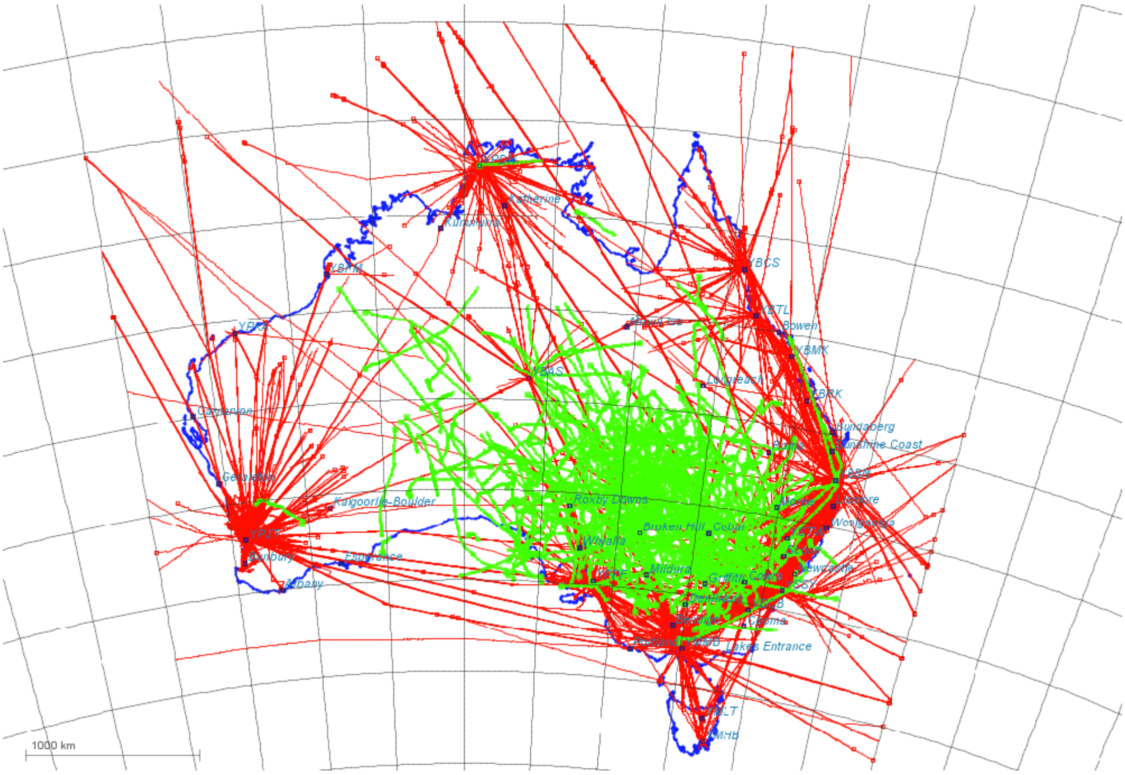

*Figure 8: Example of a random clustering of the test data*

An example of the random clusterings is shown in Figure 8. It is clear that the random clusterings are rather poor quality in that they bear no resemblance to the hand-crafted ideal.

## 3.3 External Quality Measures

First, some basic notation will be introduced. Then the external quality measures used in the experiment will be introduced.

### 3.3.1 Notation

A data set $D$ contains $N$ elements $d_1$, $d_2$, ... $d_N$.

A clustering $C$ of $D$ is a $k$-partition of $D$; that is $k$ clusters $C_1, C_2, \ldots, C_k$ such that

$$C_i \cap C_j = \phi \ \forall \ C_i, C_j \in C, C_i \neq C_j$$

and

$$\bigcup_{i=1}^{k} C_i = D$$

Let $n_i$ denote the size of cluster $C_i$.

Also let $C'$ denote a clustering, with $k'$ clusters $C'_1, C'_2, \ldots, C'_{k'}$, and $n'_j$ denote the size of cluster $C'_j$.

Then given two clusters, $C_i$ in $C$ and $C'_j$ in $C'$, let

$$m_{\mathrm{ij}} = |C_i \cap C'_j|$$

That is, $m_{ij}$ denotes the number of data elements in common between two clusters taken from different clusterings. Note that in the nomenclature of bivariate statistics, by treating the clusterings as probability distributions over their clusters, the matrix $[m_{ij}]$ would be the contingency matrix of their joint distribution.

In addition, consider the set of all unique pairs of distinct elements $(d_i, d_j)$ from $D$; then let

**M**

$\mathbf{M_{11}}$ denote the number of pairs where $d_i$ and $d_j$ are assigned to the same cluster in $C$, and $d_i$ and $d_j$ are also assigned to the same cluster under $C'$,

$\mathbf{M_{00}}$ denote the number of pairs where $d_i$ and $d_j$ are assigned to different clusters in $C$, and $d_i$ and $d_j$ are also assigned to different clusters under $C'$,

$\mathbf{M_{10}}$ denote the number of pairs where $d_i$ and $d_j$ are assigned to the same cluster in $C$, but $d_i$ and $d_j$ are assigned to different clusters under $C'$, and

$\mathbf{M_{01}}$ denote the number of pairs where $d_i$ and $d_j$ are assigned to different clusters in $C$, but $d_i$ and $d_j$ are assigned to the same cluster under $C'$.

Note that

$$M_{11} + M_{00} + M_{10} + M_{01} = \frac{N(N-1)}{2}$$

and that again these quantities can be viewed as the elements of a contingency matrix, this time between two binary variables.

Previous authors have noted that most measures described in the literature can be specified as a formula using these notations. It is the case for all the measures used in the experiment reported here. It is worth noting that these quantities are related. For example Fowlkes and Mallows (1983) note that

$$2\,M_{11} = \sum_{i=1}^{k}\sum_{j=1}^{k'} m_{\mathrm{ij}}^2 - N$$

However the implications for a potential canonical or normalised notation sufficient to describe all measures does not appear to have been explored in the literature.

### 3.3.2 Measures Used

Table 3 lists the measures used in the experiment, giving the formula of the measure and it's range. For each measure an abbreviation is given; these are used subsequently in Tables 4 and 5 to refer to the measures.

*Table 3: External Quality Measures*

| Measure | | Properties | Formula |
|---|---|---|---|
| 1 | ARI<br>Adjusted Rand Index<br>(Hubert & Arabie 1985) | Range: [-1, 1]<br>Similarity | $\frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}\binom{m_{ij}}{2}-\left[\sum_{i=1}^{k}\binom{n_i}{2}\right]\left[\sum_{j=1}^{k'}\binom{n'_j}{2}\right]/\binom{N}{2}}{\left[\sum_{i=1}^{k}\binom{n_i}{2}+\sum_{j=1}^{k'}\binom{n'_j}{2}\right]/2-\left[\sum_{i=1}^{k}\binom{n_i}{2}\right]\left[\sum_{j=1}^{k'}\binom{n'_j}{2}\right]/\binom{N}{2}}$ |
| 2 | FMI<br>Fowlkes-Mallows Index<br>(Fowlkes & Mallows 1983) | Range: [0, 1]<br>Similarity | $\frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}m_{ij}^2-N}{\sqrt{\left[\sum_{i=1}^{k}n_i^2-N\right]\left[\sum_{j=1}^{k'}n_j^2-N\right]}}$ |
| 3 | JC<br>Jaccard Coefficient | Range: [0, 1]<br>Similarity | $\frac{M_{11}}{M_{11}+M_{10}+M_{01}}$ |
| 4 | JMS<br>"Jaccard Matrix Sum"<br>(Torres, Basnet *et al* 2008) | Range: [0, 1]<br>Similarity | $\frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}\frac{m_{ij}}{\lvert C_i\cup C'_j\rvert}}{\max(k,\ k')}$ |
| 5 | MM<br>Mirkin Metric<br>(Mirkin & Cherny 1970) | Range: [0, $N^2$]<br>Distance | $2\,(M_{10}+M_{01})$ |
| 6 | NMI<br>Normalised Mutual Information<br>(Strehl & Ghosh 2002) | Range: [0, 1]<br>Distance | $\frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}\frac{m_{ij}}{N}\log\frac{\frac{m_{ij}}{N}}{\frac{n_i}{N}\frac{n'_j}{N}}}{\sqrt{\left[\sum_{i=1}^{k}\frac{n_i}{N}\log\frac{n_i}{N}\right]\left[\sum_{j=1}^{k'}\frac{n'_j}{N}\log\frac{n'_j}{N}\right]}}$ |
| 7 | RI<br>Rand Index<br>(Rand 1971) | Range: [0, 1]<br>Similarity | $\frac{M_{11}+M_{00}}{M_{11}+M_{00}+M_{10}+M_{01}}$ |
| 8 | VDM<br>van Dongen Metric<br>(van Dongen 2000) | Range: [0, 2*N*]<br>Distance | $2\,N-\sum_{i=1}^{k}\max_{j=1}^{k'}m_{ij}-\sum_{j=1}^{k'}\max_{i=1}^{k}m_{ij}$ |
| 9 | VI<br>Variation of Information<br>(Meilă 2003) | Range: [0, log*N*]<br>Distance | $\sum_{i=1}^{k}\frac{n_i}{N}\log\frac{n_i}{N}+\sum_{j=1}^{k'}\frac{n'_j}{N}\log\frac{n'_j}{N}-2\sum_{i=1}^{k}\sum_{j=1}^{k'}\frac{m_{ij}}{N}\log\frac{\frac{m_{ij}}{N}}{\frac{n_i}{N}\frac{n'_j}{N}}$ |

Notes:

- The measures described as "similarity" measures are those for which higher values are to be interpreted as the clusterings being closer or more similar. For the ones described as "distance" measures, higher values are to be interpreted as the clusterings being further apart or less similar.

- The Adjusted Rand Index is the Rand Index adjusted to be relative to the expected value with a random clustering drawn from the same marginal distribution. Ideally the adjusted index would have range [0, 1], but in practice it can adopt negative values.

- The Jaccard Coefficient measure follows naturally from the traditional measure of set similarity. It is obtained by considering the set of all unique pairs of distinct elements ($d_i$, $d_j$) from *D*. Take the subset in which both elements of the pair are assigned to the same cluster in C. Also take the subset in which both elements of the pair are assigned to the same cluster in C'. Then simply compute the Jaccard Coefficient of these two subsets.

- The title "Jaccard Matrix Sum" for the measure numbered 4 in the table above was chosen by the present author. The paper (Torres *et al* 2008) describing the measure does not give a name for it, and there do not appear to be any other extant papers citing the paper, let alone proposing a name for the measure. It is worth noting that (Torres *et al* 2008) is published in a low-quality venue and is a poor paper. It provides little analysis of the measure, and the claims regarding it are described as trivial results with no proof offered. None the less, the measure appears to be a novel and interesting approach, and is worth considering.

- The formula of the Mirkin Metric given here is taken from (Ben-Hur, Elisseeff & Guyon 2002); the metric can also be formulated as the Hamming distance between binary vector representations of clusters. The metric can also be characterised as a translation and scaling of the Rand Index (Meilă 2005).

- The Normalised Mutual Information measure is derived by treating a clustering as a probability distribution over its clusters, and then applying the well-known information-theoretic concepts of the entropy of a distribution and the mutual information of a joint distribution. In this context, if we use the typical information-theoretic notations of *H*(*C*) for the entropy of clustering *C*, *H*(*C*') for the entropy of clustering *C*', and *I*(*C*,*C*') for the mutual information of their joint distribution, then the formula for the measure is simply

  $$\frac{I(C,\, C')}{\sqrt{H(C)\, H(C')}}$$

  From this perspective, the measure is to be interpreted as quantifying the information shared by the two clusterings and normalising by the geometric mean of the individual clustering entropies (which is an upper bound on the mutual information).

- The Variation of Information measure is also derived from a straight-forward adaption of information-theoretic concepts. Using the notation above for entropy and mutual information, and also *H*(*C* | C') for conditional entropy, then the measure is simply

  $$H(C) + H(C') - 2\, I(C,\, C') = H(C \mid C') + H(C' \mid C)$$

  From this perspective, the measure is to be interpreted as quantifying the information required to describe the difference between the two clusterings.

- The measures with a range that is a function of $N$ can easily be normalised to the range [0,1] in the context of a given dataset which fixes $N$. This is convenient for some circumstances (*e.g.* comparing behaviour over multiple data sets); however in the context of the experiment reported here these normalisations are irrelevant as any scaling applied to a measure will not influence the outcome.

### 3.3.3 Measures Not Used

There are some additional measures that have been described in literature and come to the attention of this author, but which are not considered in this report due to computational tractability issues.

Meilă and Heckerman (2001) describe a measure that considers clusters as arbitrary class labels, and involves searching for the map from one clustering to the other that will minimise classification error, and using this minimised classification error as the distance between the clusterings. Zhou, Li and Zha (2005) make a similar proposal, with an alternative formulation and more complexity as soft clusterings and weights on clusters are accommodated. For both of these proposed measures, the need for solving a complex optimisation problem as part of evaluating the measures makes them unattractive. Note that the cost of the optimisation-solving portion grows super-linearly in the size of the clusterings being compared.

Vinh, Epps and Bailey (2009) describes two measures named Adjusted Mutual Information and Adjusted Variation of Information. These take Mutual Information and Variation of Information as base measures, and adjust for chance using the same adjustment formulation that Hubert and Arabie (1985) used to derive the Adjusted Rand Index. The expression they derive includes, as terms, the factorial of large numbers; e.g. the factorial of the number of elements in the data set being clustered. It was found that computing this measure using arbitrary-precision integer arithmetic on the data set and clusterings used for this experiment was intractable.

## 3.4 Discriminatory Power Evaluation

The measures listed in Table 3 were evaluated using the methodology outlined in Section 2.3 and the clusterings described in Section 3.2.

The measures were used to assess the similarity or distance between the manually constructed clustering and the algorithmically generated clusterings. The measures were also used to assess the similarity or distance between the manually constructed clustering and the randomly generated clusterings. The results are presented in Table 4 and Table 5 below. Then the formula proposed in Section 2.3 for quantifying the discriminatory power of the measures was applied. These results are presented in Table 6.

*Table 4: Similarity or distance between hand-crafted and algorithmic clusterings*

| Clustering | ARI | FMI | JC | JMS | MM | NMI. | RI | VDM | VI |
|---|---|---|---|---|---|---|---|---|---|
| Mean-30k | 0.6170 | 0.6628 | 0.4466 | 0.7208 | 29614 | 0.9579 | 0.9988 | 1368 | 0.9127 |
| Mean-40k | 0.6562 | 0.6812 | 0.4893 | 0.6882 | 49134 | 0.9445 | 0.9980 | 1549 | 1.1099 |
| Mean-50k | 0.5751 | 0.6170 | 0.4048 | 0.5713 | 69608 | 0.9282 | 0.9972 | 1909 | 1.4213 |
| Mean-60k | 0.5375 | 0.5997 | 0.3689 | 0.4960 | 86672 | 0.9168 | 0.9965 | 2071 | 1.6357 |
| Mean-70k | 0.4971 | 0.5720 | 0.3322 | 0.4318 | 103684 | 0.9046 | 0.9958 | 2255 | 1.8630 |
| Mean-80k | 0.4531 | 0.5390 | 0.2944 | 0.3838 | 124060 | 0.8935 | 0.9950 | 2418 | 2.0673 |
| PRM:50-30k | 0.7204 | 0.7245 | 0.5637 | 0.8004 | 32654 | 0.9560 | 0.9986 | 1316 | 0.8929 |
| PRM:50-40k | 0.8174 | 0.8210 | 0.6917 | 0.8164 | 17654 | 0.9694 | 0.9992 | 985 | 0.6394 |
| PRM:50-50k | 0.8180 | 0.8184 | 0.6926 | 0.8705 | 18996 | 0.9705 | 0.9992 | 911 | 0.6061 |
| PRM:50-60k | 0.7876 | 0.7908 | 0.6503 | 0.8618 | 24392 | 0.9679 | 0.9990 | 943 | 0.6526 |
| PRM:50-70k | 0.7527 | 0.7630 | 0.6043 | 0.7986 | 30944 | 0.9627 | 0.9987 | 1078 | 0.7531 |
| PRM:50-80k | 0.7279 | 0.7454 | 0.5731 | 0.7472 | 36312 | 0.9574 | 0.9985 | 1200 | 0.8556 |
| PRM:250-30k | 0.6170 | 0.6628 | 0.4466 | 0.7208 | 29614 | 0.9579 | 0.9988 | 1368 | 0.9127 |
| PRM:250-40k | 0.7727 | 0.7842 | 0.6301 | 0.7828 | 20510 | 0.9655 | 0.9991 | 1118 | 0.7301 |
| PRM:250-50k | 0.8009 | 0.8032 | 0.6684 | 0.8404 | 19602 | 0.9690 | 0.9992 | 976 | 0.6430 |
| PRM:250-60k | 0.7802 | 0.7810 | 0.6403 | 0.8656 | 23810 | 0.9667 | 0.9990 | 1010 | 0.6828 |
| PRM:250-70k | 0.7741 | 0.7776 | 0.6321 | 0.8566 | 26080 | 0.9656 | 0.9989 | 1019 | 0.7006 |
| PRM:250-80k | 0.7279 | 0.7454 | 0.5731 | 0.7472 | 36312 | 0.9574 | 0.9985 | 1200 | 0.8556 |

*Table 5: Similarity or distance between hand-crafted and random clusterings*

| Clustering | ARI | FMI | JC | JMS | MM | NMI | RI | VDM | VI |
|---|---|---|---|---|---|---|---|---|---|
| Rand-300-10 | 0.0000 | 0.0019 | 0.0009 | 0.1960 | 119572 | 0.7108 | 0.9952 | 6654 | 5.5345 |
| Rand-300-20 | 0.0004 | 0.0022 | 0.0011 | 0.2900 | 105850 | 0.7431 | 0.9957 | 6171 | 5.0276 |
| Rand-300-30 | 0.0000 | 0.0016 | 0.0008 | 0.3948 | 93726 | 0.7743 | 0.9962 | 5678 | 4.5226 |
| Rand-300-40 | 0.0005 | 0.0021 | 0.0010 | 0.4688 | 84200 | 0.7985 | 0.9966 | 5263 | 4.1144 |
| Rand-300-50 | 0.0002 | 0.0015 | 0.0006 | 0.5178 | 72778 | 0.8313 | 0.9970 | 4667 | 3.5475 |
| Rand-300-60 | 0.0003 | 0.0015 | 0.0006 | 0.5101 | 65850 | 0.8540 | 0.9973 | 4193 | 3.1421 |
| Rand-400-10 | 0.0000 | 0.0015 | 0.0007 | 0.2183 | 103210 | 0.7318 | 0.9958 | 6596 | 5.2101 |
| Rand-400-20 | 0.0000 | 0.0016 | 0.0008 | 0.3172 | 92276 | 0.7625 | 0.9963 | 6097 | 4.7164 |
| Rand-400-30 | 0.0000 | 0.0016 | 0.0007 | 0.4045 | 83932 | 0.7885 | 0.9966 | 5631 | 4.2863 |
| Rand-400-40 | 0.0000 | 0.0011 | 0.0005 | 0.5137 | 74948 | 0.8165 | 0.9970 | 5103 | 3.8089 |
| Rand-400-50 | 0.0000 | 0.0011 | 0.0004 | 0.5215 | 68328 | 0.8396 | 0.9972 | 4620 | 3.4011 |
| Rand-400-60 | 0.0000 | 0.0006 | 0.0002 | 0.5097 | 62534 | 0.8616 | 0.9975 | 4124 | 3.0053 |
| Rand-500-10 | 0.0000 | 0.0018 | 0.0009 | 0.2432 | 93586 | 0.7484 | 0.9962 | 6507 | 4.9454 |
| Rand-500-20 | 0.0006 | 0.0023 | 0.0011 | 0.3389 | 84692 | 0.7770 | 0.9966 | 5994 | 4.4772 |
| Rand-500-30 | 0.0000 | 0.0013 | 0.0006 | 0.4323 | 77084 | 0.8017 | 0.9969 | 5527 | 4.0625 |
| Rand-500-40 | 0.0006 | 0.0019 | 0.0008 | 0.5309 | 70570 | 0.8270 | 0.9971 | 4988 | 3.6241 |
| Rand-500-50 | 0.0001 | 0.0012 | 0.0004 | 0.5097 | 65040 | 0.8479 | 0.9974 | 4530 | 3.2537 |
| Rand-500-60 | 0.0000 | 0.0004 | 0.0001 | 0.5068 | 61410 | 0.8646 | 0.9975 | 4106 | 2.9509 |

*Table 6: Discriminatory power and ranking of external cluster quality measures*

| Rank | Measure | Discrim. Power | Algorithm; Mean | Algorithm; Sample Std Dev. | Random; Mean | Random; Sample Std Dev. |
|---|---|---|---|---|---|---|
| 1 | Adjusted Rand Index | 110.5 | 0.69075 | 0.11524 | 0.00003 | 0.00034 |
| 2 | Fowlkes-Mallows Index | 106.3 | 0.71611 | 0.08868 | 0.00156 | 0.00051 |
| 3 | Jaccard Coefficient | 90.10 | 0.53908 | 0.12824 | 0.00073 | 0.00028 |
| 4 | van Dongen Metric | 6.212 | 1372 | 476 | 5358 | 865 |
| 5 | Variation of Information | 5.174 | 0.9964 | 0.4474 | 4.0906 | 0.79992 |
| 6 | Normalised Mutual Information | 4.538 | 0.95068 | 0.02369 | 0.79889 | 0.04725 |
| 7 | Jaccard Matrix Sum | 2.325 | 0.72228 | 0.15222 | 0.41249 | 0.11664 |
| 8 | Rand Index | 1.703 | 0.99827 | 0.00126 | 0.99671 | 0.00066 |
|  | Mirkin Metric | 1.703 | 43314 | 31508 | 82199 | 16540 |

## 3.5 Discussion

The results in Table 6 indicate the measure with the most discriminatory power is the Adjusted Rand Index. However the second and third ranking measures - Fowlkes-Mallows Index and Jaccard Coefficient respectively - also offer a high degree of discrimination. The van Dongen Metric, and the two information-theoretic measures (Variation of Information and Normalised Mutual Information) can receive "honourable mentions" for having some discriminatory power. The three measures of Jaccard Matrix Sum, Rand Index, and Mirkin Metric perform poorly and do not exhibit any significant discriminatory power.

When the relationships between the measures are considered, these results lead to some interesting observations.

- The three best performers are all similarity measures where the empirically observed average value of the measure for random clusterings is near zero.
- The Rand Index performs the worst, but when given an adjustment for chance in the manner proposed by Hubert and Arabie, it becomes the best performer. This suggests that the adjustment is indeed favourable in practice, despite the (arguably valid) theoretical criticisms made by Meilă (2005, 2007) against the assumptions behind the adjustment.
- The result that the Rand Index and the Mirkin Metric have the same discriminatory power is to be expected. The formula for discriminatory power used here is invariant under linear translation or scaling of a measure, and as noted in Section 3.3, these two measures are related by an affine transform.
- The Jaccard Coefficient is one of the simplest measures and is the oldest measure in the sense that it is a direct application of a set similarity measure that was described in 1901. And despite being a general measure of set similarity, it performs better than most of the more complex measures specifically intended for clustering similarity that have been subsequently proposed.

# 4. Conclusion

This report addressed the problem of clustering air surveillance track data; in particular the problem of how to make a selection from amongst the numerous competing algorithms that can operate on this type of data. Section 2 discussed issues surrounding this selection task, and proposed a methodology for achieving it. Section 3 reported the experience of a trial of that methodology. The experience of this trial was promising and suggests that further use and refinement of the methodology may be a worthwhile pursuit.

Areas left for future work include:

- *Comprehensive exploration of existing or new algorithms for the quality of their output when applied to air surveillance track data.*

  The use of a test data set that has a hand-crafted ideal, and the Adjusted Rand Index for assessing how similar the output of an algorithm is to the ideal, appears to be a valid approach for selecting between competing algorithms.

- *Further investigation of external quality measures in general.*

  This is an area that has not received much attention in scholarly literature. Further analytical and empirical investigations, as seen in this report and some recent literature (Meilă 2005, 2007; Vinh, Epps & Bailey 2009), are warranted. In particular, the delineation of the general types of data and data exploitation goals that are best supported by particular measures remains open.

- *Refinement of the methodology.*

  In particular, the formula for discriminatory power warrants further attention. There is already a well-established family of statistical measures for quantifying the similarity or lack thereof between two distributions. However these statistics are typically intended to test a specific null hypothesis - for example the Kolmogorov-Smirnov statistic is used to test whether two samples are likely to have been drawn from the same distribution. This null hypothesis is not actually related to the concept of discriminatory power being used here, but none the less this family of statistical measures might offer inspiration.

# 5. References

Ben-Hur, A, Elisseeff, A & Guyon I (2002), 'A stability based method for discovering structure in clustered data', *Pacific Symposium on Biocomputing*, vol. 7.

van Dongen, S (2000), '*Performance criteria for graph clustering and Markov cluster experiments*', Technical Report INS-R0012, Centrum voor Wiskunde en Informatica.

Douglas, D & Peucker, T (1973), 'Algorithms for the reduction of the number of points required to represent a digitized line or its caricature', *The Canadian Cartographer* **10**(2).

Ester, M, Kriegel, H, Sander, J & Xu, X (1996), 'A density-based algorithm for discovering clusters in large spatial databases with noise', *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.

Fisher, R (1936), 'The use of multiple measurements in taxonomic problems', *Annual Eugenics* **7**(2).

Fowlkes, E & Mallows, C (1983), 'A method for comparing two hierarchical clusterings', *Journal of the American Statistical Association* **78**(383).

Jain, A, Murty, M & Flynn, P (1999), 'Data clustering: a review', *ACM Computing Surveys* 31(3).

Han, J & Kamber, M (2006), *Data Mining: Concepts and Techniques*', 2nd ed., Morgan Kaufmann.

Hubert, L & Arabie, P (1985), 'Comparing partitions', *Journal of Classification* 2(1).

Lee, J, Han, J & Whang, K (2007), 'Trajectory clustering: a partition-and-group framework', *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*.

MacQueen, J (1967), 'Some methods for classification and analysis of multivariate observations', *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability*, vol. 1.

Meilă, M & Heckerman, D (2001), 'An experimental comparison of model-based clustering methods', *Journal of Machine Learning* **42**(1-2).

Meilă, M (2003), 'Comparing clusterings by the variation of information', *Learning Theory and Kernel Machines (Proceedings 16th Annual Conference on Learning Theory and 7th Kernel Workshop)*, Lecture Notes in Computer Science vol. 2777.

Meilă, M (2005), 'Comparing clusterings - an axiomatic view', *Proceedings of the 22nd International Conference on Machine Learning*.

Meilă, M (2007), 'Comparing clusterings - an information based distance', *Journal of Multivariate Analysis* **98**.

Mirkin, B & Cherny, L (1970), 'Measurement of the distance between distinct partitions of a finite set of objects', *Automation and Remote Control* **31**(5).

Nguyen, Q & Rayward-Smith, V (2008), 'Internal quality measures for clustering in metric spaces', *International Journal of Business Intelligence and Data Mining* **3**(1).

Rand, W (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical Association* **66**(336).

Raskutti, B & Leckie, C (1999), 'An evaluation of criteria for measuring the quality of clusters', *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, vol 2.

Torres, G, Basnet, R, Sung, A, Mukkamala, S & Ribeiro, B (2008), 'A similarity measure for clustering and its applications', *World Academy of Science, Engineering, and Technology* **17**.

Schmitzer-Torbert, N, Jackson, J, Henze, D, Harris, K & Redish, A (2005), 'Quantitative Measures of Cluster Quality for Use in Extracellular Recordings', *Neuroscience* **131**.

Vinh, N, Epps, J & Bailey, J (2009), ''Information theoretic measure for clusterings comparison: is a correction for chance necessary?', *Proceedings of the 26th International Conference on Machine Learning*.

Zhou, D, Lai, J & Zha, H (2005), 'A new Mallows distance based metric for comparing clusterings', *Proceedings of the 22nd International Conference on Machine Learning*.

Zhang, T, Ramakrishnan, R & Livny, M (1996), 'BIRCH: an efficient data clustering method for very large databases', *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*.

Page classification: UNCLASSIFIED

| | |
|---|---|
| **DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION**<br>**DOCUMENT CONTROL DATA** | <br>1. PRIVACY MARKING/CAVEAT (OF DOCUMENT) |

| | |
|---|---|
| 2. TITLE<br>Towards an Evaluation of Air Surveillance Track Clustering Algorithms via External Cluster Quality Measures | 3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)<br>Document (U)<br>Title (U)<br>Abstract (U) |
| 4. AUTHOR(S)<br>Matthew C. Lowry | 5. CORPORATE AUTHOR<br>Defence Science and Technology Organisation<br>PO Box 1500<br>Edinburgh South Australia 5111 Australia |

| | | | |
|---|---|---|---|
| 6a. DSTO NUMBER<br>DSTO-TR-2800 | 6b. AR NUMBER<br>AR-015-522 | 6c. TYPE OF REPORT<br>Technical Report | 7. DOCUMENT DATE<br>January 2013 |

| | | | | |
|---|---|---|---|---|
| 8. FILE NUMBER<br>2012/1237568/1 | 9. TASK NUMBER<br>INT07/356 | 10. TASK SPONSOR<br>DIO | 11. NO. OF PAGES<br>26 | 12. NO. OF REFERENCES<br>24 |

| | |
|---|---|
| 13. DSTO Publications Repository<br>http://dspace.dsto.defence.gov.au/dspace/ | 14. RELEASE AUTHORITY<br>Chief, Command, Control, Communications and Intelligence Division |

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

*Approved for public release*

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS

Algorithms, Data Mining, Cluster analysis, Air surveillance

19. ABSTRACT

Clustering is a data mining technique for analysing large data sets and finding groups of elements within the data set that are similar to each other. The use of clustering on archives of historical air surveillance track data would enable the discovery of flights that exhibited similar behaviour and followed similar flight paths. However there are many different clustering algorithms available, so some method for selecting the best from the competing algorithms is required. Unfortunately the academic literature has yet to provide a general, comprehensive, and robust methodology for this task. Further the niche nature of the problem domain means the academic literature provides no direct assistance by way of reporting practical experience in the use of particular algorithms on air surveillance track data. This report aims to fill the gap by describing such a methodology for evaluating and choosing between competing clustering algorithms.

Page classification: UNCLASSIFIED